IDIAP RESEARCH REPORT

THE SIWIS DATABASE: A MULTILINGUAL SPEECH DATABASE WITH ACTED EMPHASIS

Jean-Philippe Goldman Pierre-Edouard Honnet[a]
Rob Clark Philip N. Garner Maria Ivanova
Alexandros Lazaridis[a] Hui Liang Tiago Macedo
Beat Pfister Manuel Sam Ribeiro Eric Wehrli
Junichi Yamagishi

Idiap-RR-13-2016

MAY 2016

[a]Idiap Research Institute

Centre du Parc, Rue Marconi 19, P.O. Box 592, CH - 1920 Martigny
T +41 27 721 77 11 F +41 27 721 77 12 info@idiap.ch www.idiap.ch

The SIWIS database: a multilingual speech database with acted emphasis

Jean-Philippe Goldman, Pierre-Edouard Honnet, Rob Clark, Philip N. Garner, Maria Ivanova, Alexandros Lazaridis, Hui Liang, Tiago Macedo, Beat Pfister, Manuel Sam Ribeiro, Eric Wehrli and Junichi Yamagishi

April 5, 2016

Abstract

We describe here a collection of speech data of bilingual and trilingual speakers of English, French, German and Italian. In the context of speech to speech translation (S2ST), this database is designed for several purposes and studies: training CLSA systems (cross-language speaker adaptation), conveying emphasis through S2ST systems, and evaluating TTS systems. More precisely, 36 speakers judged as accentless (22 bilingual and 14 trilingual speakers) were recorded for a set of 171 prompts in two or three languages, amounting to a total of 24 hours of speech. These sets of prompts include 100 sentences from news, 25 sentences from Europarl, the same 25 sentences with one acted emphasised word, 20 semantically unpredictable sentences, and finally a 240-word long text. All in all, it yielded 64 bilingual session pairs of the six possible combinations of the four languages. The database is freely available for non-commercial use and scientific research purposes.

Index Terms: speech-to-speech translation, speech corpus, bilingual speakers, emphasis

1 Introduction

In the context of speech-to-speech translation (S2ST), the SIWIS research project[1] is a Swiss-NSF-funded project gathering several research teams in Switzerland and the CSTR (University of Edinburgh) [1]. It was inspired by the EMIME project [2] in which languages such as English, Japanese, Mandarin and Finnish, were involved. For SIWIS, we focused on the three main official languages in Switzerland (French, German, Italian) and English. Besides, an additional purpose of SIWIS is an attempt of conveying speaker intents through prosody.

Tsiartas et al. [3] showed in a large scale human evaluation framework that the perceived quality of S2ST was correlated with cross-lingual prosodic emphatic transfer. In other words, emphasising the correct words in the output language in TTS based on the emphasised words in the input language helps in the S2ST task. This observation motivates the need for emphasised data in our bilingual corpus, as parallel sentences in both emphasised and neutral version can be used for emphasis translation.

In this sense, the speech corpus should be useful for emphasis analysis, cross-lingual emphasis and intent transfer, cross-lingual adaptation with parallel speech from same speakers, cross-lingual studies in general. It has already been exploited successfully for emphasis detection evaluation [4, 5]. The bilingual aspect of the database also enabled investigation on speakers' prosody when they speak different languages[6].

The EMIME speech database was used as a basis for the design of the SIWIS database [7]. We recorded 36 accentless bilingual speakers (among which 14 trilingual ones) yielding 86 bilngual pairs of set of 171 prompts in two of the four languages, i.e. almost 24 hours of speech. The reading material is mainly composed of news or parliamentary sentences. Besides, some sentences were repeated with some emphasis. Additionnaly, a 240-word text was read with some involvment.

In its second section, this contribution describes how the speakers were selected, whereas the third section gives details on how they were recorded on the reading material. Eventually, the fourth section shows additional annotations and processings such as labelling and alignment.

[1] `https://www.idiap.ch/project/siwis`

2 Speaker selection

All the speakers were selected on the basis of small recordings that could be done over the Internet on a dedicated webpage (`http://bit.ly/bilinguals`). Advertisement for this task was done through ads, flyers and mailing-lists within academic institutes, mainly Swiss universities (Geneva, Neuchâtel, Zurich) and international non-governmental organisations in Geneva.

On this webpage, the candidates were asked for their e-mail, age and for each language they would apply for, their A–B–C level, at which age they started this language, and if a regional accent could be perceptible, even slightly[2]. For each of the applied language, the candidates could be recorded as they were reading a short excerpt of *"Le Petit Prince / The Little Prince"* of Antoine de Saint-Éxupéry. The passages in all 4 languages taken for the website (`http://bit.ly/petit_prince`), showing this novel in 100+ languages, were 70 to 75 words in length.

All candidates answering to all information and having applied and recorded their voice in at least two languages were pre-selected and their recordings were sent to 3 (sometimes 4) native judges of each language. The judges were expert in linguistics and were asked to evaluate candidate for their accentedness in the different languages on a 0-3 scale with possibilities to add comments.

- 0 = strong foreign accent
- 1 = noticeable foreign accent
- 2 = very slight foreign accent
- 3 = no foreign accent

Discarding incomplete application and candidates with only one recording, a total of 137 candidates were registered. Their age was 26 in average (s.d. 10 yrs) with a minimum at 10 and a maximum at 89. Most of them applied for 2 languages (91 bilingual speakers), about one third as trilingual speakers (39 candidates) and only 7 quadrilingual speakers. Table 1 shows for each recording the A-B-C level pretended by the candidates.

Table 1: *Total recordings per language (and % claiming to be A B C).*

	French	English	German	Italian
Total	118	110	52	47
%A	69	39	65	78
%B	28	56	20	19
%C	4	5	5	3

After evaluation by native judges, only a fraction of candidates were selected as speakers. The main rule was to select candidates with an average evaluation of 2.5 at least and with no evaluation below 2. In short, most of the speakers were evaluated with no foreign accents by all three judges (three '3's). A small proportion was evaluated with a slight foreign accent by one judge whereas the two others have evaluated him with no foreign accent (one '2' and two '3's). Some trilingual and quadrilingual candidate failing to have the required evaluation in one language, have then been selected for a lower number of languages. Table 2 depicts the average evaluation of speakers for each language.

Table 2: *Average evaluation of candidates for each language. Scale used by the judges: 0 = strong foreign accent, 1 = noticeable accent, 2 = very slight accent, 3 = no foreign accent.*

	French	English	German	Italian
$= 3$	55	16	11	4
$\geqslant 2.5$	13	5	14	11
$\geqslant 2$	11	24	11	16
< 2	39	65	14	16
Total	118	110	52	47

All in all, 36 speakers could effectively be recorded with 22 bilingual and 14 trilingual speakers. The 22 bilingual speakers were recorded in 2 languages, yielding 44 recording sessions, and the 14 trilingual

[2]A–B–C language level is generally used by translators and interpreters to denote respectively A as their main language, usually mother tongue, B as another active language of which they have an excellent command, and C as a passive language, which is used only as a source language for translation and interpretation.

speakers were record in 3 languages, yielding 42 recording sessions. Table 3 shows the number of bilingual and trilingual speakers by genre. Details on how the recording sessions occurred as well as the reading material are explained in the next session.

Table 3: *Number of bilingual and trilingual speakers by genre among the 36 speakers.*

	Bilingual	Trilingual	Total
Female	10	11	21
Male	12	3	15
Total	22	14	36

The 86 recording sessions were combined accordingly to the wanted pair of languages into 63 pairs of recording sessions. The table below shows to the number of pairs of recording sessions per language.

Table 4: *Number of pairs of recording sessions per language.*

Language pair	Number of session pairs (male + female)
French-English	20 (9 + 11)
French-German	12 (5 + 7)
French-Italian	13 (6 + 7)
English-German	10 (3 + 7)
English-Italian	5 (0 + 5)
German-Italian	4 (1 + 3)
Total	64 (24 + 40)

3 Recordings

This section describes the recording sessions *per se*. The selected bilingual speakers were paid CHF 60.- (and 90.- for trilingual speakers) and had to sign an informed consent. Each recording session (i.e. all the prompts in one language) took about 20 minutes and speakers could make a large pause between the two or three sessions. As the task could be exhausting, the weakest language was generally done first.

3.1 Reading material

The stimulus material was largely inspired by the EMIME bilingual corpus [7] to keep consistency and allow future studies involving both corpus[3]. In our case, each set of 171 prompts for each language is divided in 5 parts as follows:

- EUROPARL (prompts numbered as 000 to 024): 25 Europarl statements among which 20 declaratives and 5 interrogatives. The Europarl corpus was used to have a parallel meaning across languages.
- NEWS (100-199): 100 sentences from newspapers among which 80 declaratives and 20 interrogatives.
- SUS (200-219): 20 SUS, or semantically unpredictable sentences. e.g sentence #200 (of scenario A)
  - Le chien lutte sous la plage rouge.
  - The dog fights under the red beach.
  - Das Haar steht auf dem leichten Zahn.
  - Il cane lotta contro la spiaggia rossa
- FOCUS (300-324): 25 Europarl statements. These are the same as in part EUROPARL but one word, written in capital in the prompt, is emphasised, i.e pronounced with a focus. e.g.
  - Je VOIS ce que vous voulez dire
  - I SEE what you are saying.
  - Ich VERSTEHE, was Sie meinen.

[3]The reading material of EMIME consists in 25 Europarl sentences, 100 news sentences and 20 semantically unpredictable sentences (SUS)

- CAPISCO quello che intende dire.

- PRINCE (400): Text reading “Le petit prince”. The selected continuous passage has a length of about 240 words with some interrogative sentences and some direct and indirect discourse. The text was presented as a single prompt to ensure consistency in the prosody. The speaker was asked to read it with involvement.

As a reminder, EUROPARL, SUS, FOCUS and PRINCE parts have a parallel meaning across the 4 languages. Moreover, to insure variety in the uttered prompts, each language has 3 scenarii named A, B and C. In other words, each language has 3 differents sets of prompts (keeping the parallel meaning across language within each scenario). Only the 5th part (PRINCE) is the same for all the speakers.

3.2 Hardware and software

The recordings took place in a anechoic booth in which a dynamic microphone SHURE MX418/C at 10-20 cm from the speaker with a pop shield, and a keyboard to control the prompts scrolling were placed. The prompts were visible to the subject on a screen outside of the booth. A clone screen was visible to the operator to supervise the session. The sound device USBPre2 was used to record the signal into a 44.1kHz-mono-16bits format.

The SpeechRecorder[4] software (from the Institute of Phonetics and Speech Processing of the Ludwig-Maximilians-Universität München) was used to present the prompts one by one. The prompts were randomly mixed within the 4 first parts (i.e excepted from the PRINCE part which was presented as a unique prompt). The speaker was presented the prompt on the screen, could take a few seconds to read it mentally, then pronounced it and had to press a key to either jump to the next prompt or re-record the same prompt. Redoing the same prompt was done in case of stuttering, hesitation or wrong reading. The speaker usually realised he had to restart the same prompt by himself. Nevertheless, the operator could also ask the speaker to do so.

3.3 Statistics on recordings

Table 5 sums up the number of sessions, sound files (prompts) and total duration per language.

Table 5: *Recording numbers and durations.*

Language	Sessions	Prompts	Total duration
French	31	5332	512 min.
English	22	3771	350 min.
German	17	2903	266 min.
Italian	16	2738	287 min.
Total	86	14744	1415 min. ~ 23.6 hrs.

4 Additional annotation

In addition to the audio recordings and corresponding transcriptions, we created labels that can be used for statistical parametric speech synthesis, or for speech analysis.

Label format

The labels were created to the HTS [8] full context format for three of the four languages: English, French and German. It consists of linguistic features at the phone, syllable, word, phrase, sentence levels, with information such as stress, accent, part-of-speech (for details, see the file *lab_format.pdf* in the HTS demo[5]). To create the labels, we used two different text analysis front end softwares: Festival for English and German [9], and eLite for French [10].

Alignment

The labels were forced aligned using Viterbi algorithm. We used HMM-based speech synthesis models to estimate the alignment of the labels from the audio. Our models were trained using speaker adaptive training [11]. For English, the models were trained on the Wall Street Journal database [12]; for German, we used PhonDat [13], and for French, we trained our models on BREF [14]. Almost all the English, French and German labels were forced aligned. No manual correction were done on the labels. The

[4] `www.bas.uni-muenchen.de/Bas/software/speechrecorder/`

[5] Available at `http://hts.sp.nitech.ac.jp/?Download`

resulting labels provide alignment at the phone level and state level (where the states correspond to HMM states with standard settings).

Augmenting the labels with emphasis information

As part of the database contains acted emphasis, some of the labels were augmented with emphasis labels. In addition to the standard contextual features, we added a binary feature that corresponds to the question *"is the current word emphasised?"*. This additional feature was manually annotated on the labels aligned at the phone level for English, French and German, on the subsets A and C of the sentences containing emphasis. This additional information, together with the forced alignment, can be used for easy analyses of emphasised segment, or for training or adapting models which discriminate emphasis.

Current status of the annotations

Table 6 provides the number of files for which label exist, and the number of aligned labels which have emphasis marks per language.

Table 6: *Labels and emphasis.*

Language	Aligned labels	With emphasis marks
French	4474	440
English	3597	303
German	2561	276
Italian	X	X
Total	10632	1019

We plan to create labels for Italian data, and to align these in a similar fashion as for English, French and German. Some missing labels in the other languages also need to be aligned. Another task to be completed is the annotation of emphasis for all the sentences which comprise explicitely emphasised words.

5 Conclusion

This paper presented a speech database containing parallel speech recordings of bilingual and trilingual speakers in the official Swiss languages (French, German and Italian), as well as English. Another feature of the corpus is the word level emphasis acted by the speakers, in a parallel manner – both neutral and emphasised version of the sentences are available. The data presented will thus enable studies on multilingual systems as well as on emphasis in a S2ST context. Some research has already been performed succesfully using various aspects of the database.

Further refinements to this speech database include additional recordings to balance the language pairs. The creation and alignment of all the labels should also be performed, as well as the annotation of emphasis on the relevant files.

The database is freely available for non-commercial use and scientific research purposes at `http://bit.ly/siwisData`.

6 Acknowledgements

This work has been conducted with the support of the Swiss NSF under grant CRSII2 141903: Spoken Interaction with Interpretation in Switzerland (SIWIS).

References

[1] P. N. Garner, R. Clark, J.-P. Goldman, P.-E. Honnet, M. Ivanova, A. Lazaridis, H. Liang, B. Pfister, M. S. Ribeiro, E. Wehrli, and J. Yamagishi, "Translation and prosody in Swiss languages," in *Nouveaux cahiers de linguistique française*, 2014.

[2] M. Kurimo, W. Byrne, J. Dines, P. N. Garner, M. Gibson, Y. Guan, T. Hirsimäki, R. Karhila, S. King, H. Liang, K. Oura, L. Saheer, M. Shannon, S. Shiota, J. Tian, K. Tokuda, M. Wester, Y.-J. Wu, and J. Yamagishi, "Personalising speech-to-speech translation in the EMIME project," in *Proceedings of the ACL 2010 System Demonstrations*, Uppsala, July 2010, pp. 48–53.

[3] A. Tsiartas, P. G. Georgiou, and S. S. Narayanan, "A study on the effect of prosodic emphasis transfer on overall speech translation quality," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Vancouver, Canada: IEEE, 2013.

[4] M. Cernak and P.-E. Honnet, "An empirical model of emphatic word detection," in *Proceedings of Interspeech*, Dresden, Germany, September 2015.

[5] H. Liang, "Detecting emphasised spoken words by considering them prosodic outliers and taking advantage of HMM-based TTS framework," in *Speech Prosody Conference*, Boston, USA, 2016.

[6] J.-P. Goldman and S. Schwab, "Do speakers show different F_0 when they speak in different languages? The case of English, French and German," in *Speech Prosody Conference*, Boston, USA, 2016.

[7] M. Wester, "The EMIME bilingual database," The University of Edinburgh, Tech. Rep., 2010.

[8] H. Zen, T. Nose, J. Yamagishi, S. Sako, T. Masuko, A. Black, and K. Tokuda, "The HMM-based speech synthesis system (HTS) version 2.0," in *Proceedings of the 6th ISCA Speech Synthesis Workshop*, 2007, pp. 294–299.

[9] A. Black, P. Taylor, and R. Caley, "The Festival Speech Synthesis System," Human Communication Research Centre, University of Edinburgh, Technical Report, 1997.

[10] S. Roekhaut, S. Brognaux, R. Beaufort, and T. Dutoit, "eLite-HTS: A NLP tool for French HMM-based speech synthesis," in *Proceedings of Interspeech*, 2014, pp. 2136–2137.

[11] J. Yamagishi, T. Kobayashi, Y. Nakano, K. Ogata, and J. Isogai, "Analysis of speaker adaptation algorithms for HMM-based speech synthesis and a constrained SMAPLR adaptation algorithm," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 17, no. 1, pp. 66–83, 2009.

[12] D. B. Paul and J. M. Baker, "The design for the wall street journal-based csr corpus," in *Proceedings of the Workshop on Speech and Natural Language*, Stroudsburg, PA, USA, 1992, pp. 357–362.

[13] W. J. Hess, K. J. Kohler, and H.-G. Tillmann, "The Phondat-verbmobil speech corpus." in *Proceedings of EUROSPEECH*, 1995.

[14] L. F. Lamel, J.-L. Gauvain, and M. Eskenazi, "BREF, a large vocabulary spoken corpus for French," in *Proceedings of EUROSPEECH*, 1991, pp. 505–508.